

Bài 4: KHAI THÁC MẪU KẾT HỢP

I. Mục tiêu:

Sau khi thực hành xong, sinh viên nắm được:

- Thuật toán Apriori
- Áp dụng thuật toán này

II. Tóm tắt lý thuyết:

1. Các khái niệm cơ bản:

Cho các item $I = i_1, \dots, i_m$ và cơ sở dữ liệu giao dịch $D = t_1, \dots, t_n$

TID	Itemset
t_1	i_1, i_2, i_m
t_2	i_1
...	
t_n	i_2, i_m

hoặc

	i_1	i_2	...	i_m
t_1	1	1	...	1
t_2	1	0	...	0
\vdots	\vdots	\vdots	\vdots	\vdots
t_n	0	1	...	1

- Hạng mục (item)
- Tập các hạng mục (itemset)
- Giao dịch (transaction)
- Mẫu phổ biến (frequent item)
- Tập k-hạng mục (k-itemset)
- Độ phổ biến (support): được tính bằng số các giao dịch chứa item X chia cho tổng số giao dịch và được tính bằng công thức: $supp(X) = \frac{count(X)}{|D|}$.
- Tập phổ biến (frequent itemset): là tập các hạng mục S (itemset) thỏa mãn độ phổ biến tối thiểu (minsupp – do người dùng xác định). Nếu $supp(S) \geq minsupp$ thì S là tập phổ biến.
- Tập phổ biến tối đại (max pattern) thỏa
 - $supp(X) \geq minsupp$
 - Không tồn tại $|X'| > |X|$ với X' cũng phổ biến
- Tập phổ biến đóng (closed pattern)
 - $supp(X) \geq minsupp$

- Không tồn tại $|X'| > |X|$ mà $supp(X') = supp(X)$
- Luật kết hợp (association rule): ký hiệu $X \rightarrow Y$, nghĩa là khi X có mặt thì Y cũng có mặt (với xác suất nào đó)
- Độ tin cậy (confidence): được tính bằng công thức $conf(X \rightarrow Y) = \frac{supp(X,Y)}{supp(X)}$

2. Thuật toán Apriori:

Thuật toán Apriori bắt đầu bằng việc đếm các support của các item riêng biệt để khởi tạo 1-itemsets phổ biến. Tập 1-itemsets được kết hợp để tạo ra 2-itemset ứng cử viên mà support của nó được đếm. Tập 2-itemsets được tiếp tục dùng. Tổng quát, các itemset chiều dài k được sử dụng để khởi tạo các ứng cử viên chiều dài (k + 1) cho việc tăng giá trị của k. Cho \mathcal{F}_k ký hiệu tập hợp k-itemsets phổ biến, và \mathcal{C}_k ký hiệu tập hợp k-itemsets ứng cử. Lỗi của xấp xỉ là để khởi tạo lặp lại (k + 1)-ứng cử viên \mathcal{C}_{k+1} từ k-itemsets trong \mathcal{F}_k đã được tìm thấy bởi thuật toán. Các mẫu thường xuyên của (k+1)-ứng cử này được tính đối với cơ sở dữ liệu giao dịch (transaction). Trong khi việc khởi tạo (k+1)-ứng cử viên, không gian tìm kiếm có thể được xén bớt bởi việc kiểm tra tất cả k-subset của \mathcal{C}_{k+1} hay không được bao gồm trong \mathcal{F}_k .

Nếu 1 cặp itemset X và Y trong \mathcal{F}_k có (k-1) item chung thì sự kết nối giữa chúng sử dụng (k-1) item chung sẽ khởi tạo một itemset ứng cử kích thước (k+1). Ví dụ, 2 tập 3-itemset {a, b, c} (hoặc abc cho ngắn gọn) và {a, b, d} (hoặc abd cho ngắn gọn), khi chúng kết nối với nhau trong 2 item chung a và b, sẽ sinh ra 4-itemset ứng cử abcd.

Thuật toán Apriori được phát biểu như sau:

```

Algorithm Apriori(Transactions:  $\mathcal{T}$ , Minimum Support: minsup)
begin
   $k = 1$ ;
   $\mathcal{F}_1 = \{ \text{All Frequent 1-itemsets} \}$ ;
  while  $\mathcal{F}_k$  is not empty do begin
    Generate  $\mathcal{C}_{k+1}$  by joining itemset-pairs in  $\mathcal{F}_k$ ;
    Prune itemsets from  $\mathcal{C}_{k+1}$  that violate downward closure;
    Determine  $\mathcal{F}_{k+1}$  by support counting on  $(\mathcal{C}_{k+1}, \mathcal{T})$  and retaining
      itemsets from  $\mathcal{C}_{k+1}$  with support at least minsup;
     $k = k + 1$ ;
  end;
  return  $(\cup_{i=1}^k \mathcal{F}_i)$ ;
end

```

Ví dụ: Cho tập dữ liệu gồm 6 giao dịch với 0 biểu diễn sự vắng mặt của một item và 1 biểu diễn sự có mặt của nó

Transaction ID	Wine	Chips	Bread	Milk
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

Cho $\text{minsup} = 50\%$, $\text{min confidence} = 80\%$.

- $k = 1$, khởi tạo \mathcal{F}_1

Item	frequency
Wine	4
Chips	4
Bread	4
Milk	5

- Khởi tạo \mathcal{C}_2 bằng việc kết hợp các cặp item của \mathcal{F}_1
 $\{\{\text{Wine, Chips}\}, \{\text{Wine, Bread}\}, \{\text{Wine, Milk}\}, \{\text{Chips, Bread}\}, \{\text{Chips, Milk}\}, \{\text{Bread, Milk}\}\}$
- Tạo \mathcal{F}_2

Item	frequency
Wine, Chips	3
Wine, Bread	3
Wine, Milk	4
Chips, Bread	2
Chips, Milk	3
Bread, Milk	4

- Tìm các hạng mục quan trọng dựa vào minsup = 50% \Rightarrow chỉ lấy các 2-item sau:
 $\{\text{Wine, Milk}\}, \{\text{Bread, Milk}\}$
- Phát sinh các luật

$$\text{Wine} \rightarrow \text{Milk} \text{ có } \text{conf}(\text{Wine} \rightarrow \text{Milk}) = \frac{\text{support}(\text{Wine, Milk})}{\text{support}(\text{Wine})} = \frac{\frac{4}{6}}{\frac{4}{6}} = 100\%$$

$$\text{Milk} \rightarrow \text{Wine} \text{ có } \text{conf}(\text{Milk} \rightarrow \text{Wine}) = \frac{\text{support}(\text{Wine, Milk})}{\text{support}(\text{Milk})} = \frac{\frac{4}{6}}{\frac{5}{6}} = 80\%$$

$$\text{Bread} \rightarrow \text{Milk} \text{ có } \text{conf}(\text{Bread} \rightarrow \text{Milk}) = \frac{\text{support}(\text{Bread, Milk})}{\text{support}(\text{Bread})} = \frac{\frac{4}{6}}{\frac{4}{6}} = 100\%$$

$$\text{Milk} \rightarrow \text{Bread} \text{ có } \text{conf}(\text{Milk} \rightarrow \text{Bread}) = \frac{\text{support}(\text{Milk, Bread})}{\text{support}(\text{Milk})} = \frac{\frac{4}{6}}{\frac{5}{6}} = 80\%$$

- Ở bước lược bỏ, ta có $\mathcal{F}_2 = \{(\text{Wine, Milk}), (\text{Bread, Milk})\}$
- Ở bước kết hợp các item của \mathcal{F}_2 , ta có \mathcal{C}_3 gồm cặp 3-item là $\{\text{Wine, Bread, Milk}\}$
- Tạo \mathcal{F}_3

Itemset	Frequency
Wine, Bread, Milk	3

- Tìm các hạng mục quan trọng dựa vào minsup = 50% $\Rightarrow \mathcal{F}_3 = \emptyset$
 \Rightarrow Thuật toán kết thúc

III. Nội dung thực hành:

1. Cài đặt thuật toán Apriori

- Cho CSDL với các giao dịch sau:

TID	Itemset
1	Wine, Chips, Bread, Butter, Milk, Apple
2	Wine, Bread, Butter, Milk
3	Bread, Butter, Milk
4	Chips, Apple
5	Wine, Chips, Bread, Butter, Milk, Apple
6	Wine, Chips, Milk
7	Wine, Chips, Bread, Butter, Apple
8	Wine, Chips, Milk
9	Wine, Bread, Apple
10	Wine, Bread, Butter, Milk
11	Chips, Bread, Butter, Apple
12	Wine, Butter, Milk, Apple
13	Wine, Chips, Bread, Butter, Milk
14	Wine, Bread, Milk, Apple
15	Wine, Bread, Butter, Milk, Apple
16	Wine, Chips, Bread, Butter, Milk, Apple
17	Chips, Bread, Butter, Milk, Apple
18	Chips, Butter, Milk, Apple
19	Wine, Chips, Bread, Butter, Milk, Apple
20	Wine, Bread, Butter, Milk, Apple
21	Wine, Chips, Bread, Milk, Apple
22	Chips

- Tạo file “data.csv” như trong hình

	A	B	C	D	E	F
1	Wine	Chips	Bread	Butter	Milk	Apple
2	Wine		Bread	Butter	Milk	
3			Bread	Butter	Milk	
4		Chips				Apple
5	Wine	Chips	Bread	Butter	Milk	Apple
6	Wine	Chips			Milk	
7	Wine	Chips	Bread	Butter		Apple
8	Wine	Chips			Milk	
9	Wine		Bread			Apple
10	Wine		Bread	Butter	Milk	
11		Chips	Bread	Butter		Apple
12	Wine			Butter	Milk	Apple
13	Wine	Chips	Bread	Butter	Milk	
14	Wine		Bread		Milk	Apple
15	Wine		Bread	Butter	Milk	Apple
16	Wine	Chips	Bread	Butter	Milk	Apple
17		Chips	Bread	Butter	Milk	Apple
18		Chips		Butter	Milk	Apple
19	Wine	Chips	Bread	Butter	Milk	Apple
20	Wine		Bread	Butter	Milk	Apple
21	Wine	Chips	Bread		Milk	Apple
22		Chips				

- Cài đặt apyori: pip install apyori

```

C:\WINDOWS\system32\cmd.exe
Microsoft Windows [Version 10.0.19045.2846]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Huynh>pip install apyori
Collecting apyori
  Downloading apyori-1.1.2.tar.gz (8.6 kB)
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: apyori
  Building wheel for apyori (setup.py) ... done
  Created wheel for apyori: filename=apyori-1.1.2-py3-none-any.whl size=5980 sha256=423b96694217c25d3e265a2789353b9d4cf78ed3113b210dc8fc613cc962046e
  Stored in directory: c:\users\huynh\appdata\local\pip\cache\wheels\cb\xf6\xe1\57973c631d27efd1a2f375bd6a83b2a616c4021f24aab84080
Successfully built apyori
Installing collected packages: apyori
Successfully installed apyori-1.1.2

```

- Import các thư viện và load dữ liệu

```
>>> import numpy as np
>>> import pandas as pd
>>> from apyori import apriori
>>> data = pd.read_csv('D:\\Huynh\\DataMining_Lab\\data\\tuan4\\data.csv', header=None)
>>> data
```

	0	1	2	3	4	5
0	Wine	Chips	Bread	Butter	Milk	Apple
1	Wine	NaN	Bread	Butter	Milk	NaN
2	NaN	NaN	Bread	Butter	Milk	NaN
3	NaN	Chips	NaN	NaN	NaN	Apple
4	Wine	Chips	Bread	Butter	Milk	Apple
5	Wine	Chips	NaN	NaN	Milk	NaN
6	Wine	Chips	Bread	Butter	NaN	Apple
7	Wine	Chips	NaN	NaN	Milk	NaN
8	Wine	NaN	Bread	NaN	NaN	Apple
9	Wine	NaN	Bread	Butter	Milk	NaN
10	NaN	Chips	Bread	Butter	NaN	Apple
11	Wine	NaN	NaN	Butter	Milk	Apple
12	Wine	Chips	Bread	Butter	Milk	NaN
13	Wine	NaN	Bread	NaN	Milk	Apple
14	Wine	NaN	Bread	Butter	Milk	Apple
15	Wine	Chips	Bread	Butter	Milk	Apple
16	NaN	Chips	Bread	Butter	Milk	Apple
17	NaN	Chips	NaN	Butter	Milk	Apple
18	Wine	Chips	Bread	Butter	Milk	Apple
19	Wine	NaN	Bread	Butter	Milk	Apple
20	Wine	Chips	Bread	NaN	Milk	Apple
21	NaN	Chips	NaN	NaN	NaN	NaN

- Chuyển DataFrame thành dạng danh sách (list)

```
>>> records = []
>>> for i in range(0,22):
    records.append([str(data.values[i,j]) for j in range(0,6)])
```

- Xây dựng mô hình Apriori

```
>>> association_rules = apriori(records, min_support=0.50, min_confidence=0.7, min_lift=1.2,min_length=2)
>>> association_results = list(association_rules)
```

- In kết quả

```
>>> def inspect(output):
    lhs = [tuple(result[2][0][0])[0] for result in output]
    rhs = [tuple(result[2][0][1])[0] for result in output]
    support = [result[1] for result in output]
    confidence = [result[2][0][2] for result in output]
    lift = [result[2][0][3] for result in output]
    return list(zip(lhs, rhs, support, confidence, lift))

>>> output_DataFrame = pd.DataFrame(inspect(association_results), columns = ['Left_Hand_Side', 'Right_Hand_Side', 'Support', 'Confidence', 'Lift'])
>>> output_DataFrame
```

	Left_Hand_Side	Right_Hand_Side	Support	Confidence	Lift
0	Butter	Bread	0.5	0.733333	1.241026

2. Yêu cầu:

- Cài đặt thuật toán Apriori
- Viết file báo cáo trình bày tóm tắt code em đã làm và so sánh kết quả giữa hàm em viết với hàm có sẵn trong Python.